SOFTWAREPARK HAGENBERG - AUS IDEEN WERDEN ERFOLGE

Big Data Analysis in Life Sciences: Cloud Computing für die Genomanalyse

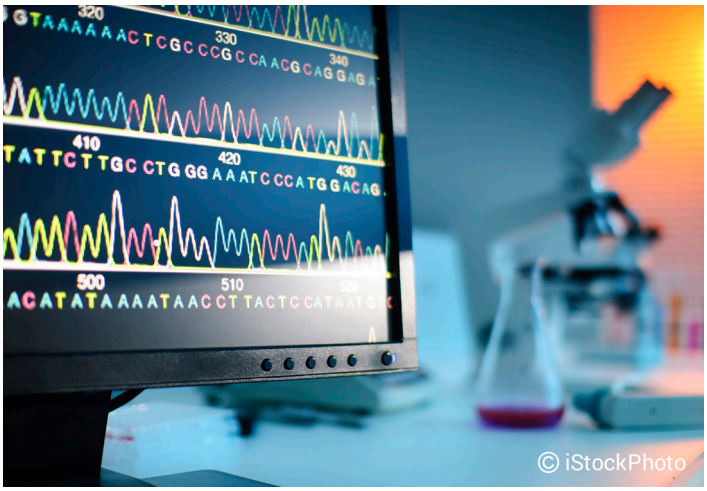
Die Analyse großer Datenmengen hat in den letzten Jahren in Wirtschaft und Forschung signifikant an Bedeutung gewonnen. Technologische Fortschritte ermöglichen es in allen Bereichen immer größere Datenmengen zu generieren und zu sammeln.

Besonders im Bereich der Life Sciences entstehen mittels modernster High-Throughput-Methoden wie bei der DNS-Sequenzierung große Mengen an biologischen Daten. Mittlerweile ist es möglich, menschliche Genome in weniger als einem Tag zu sequenzieren. Diese großen Datenmengen stellen einerseits eine technologische Herausforderung, andererseits aber eine Chance zur Gewinnung wertvoller Informationen dar.

Da traditionelle Bioinformatikanwendungen oft für die lokale Verarbeitung von Daten konzipiert sind, jedoch lokale Rechner nicht immer die gewünschte Leistung aufweisen, hat die RISC Software GmbH Möglichkeiten zum Einsatz von High Performance Computing, Grid Computing und Cloud Computing von der effizienten Verarbeitung bis hin zur interaktiven Visualisierung von Sequenzdaten analysiert.

Anwendungsfall: EU-Projekt Mr.SymBioMath

Hauptziel des von der EU ab Februar 2013 im Rahmen des Marie Curie Programms „Industry-Academia Partnerships and Pathways“ geförderten Projekts Mr.SymBioMath (http://mrsymbiomath.eu/) sind Arbeiten aus dem Bereich der vergleichenden Genomik basierend auf Cloud Computing und Hochleistungsrechnen. Die RISC Software GmbH ist dabei Teil eines internationalen Konsortiums, dem des Weiteren die Universität Malaga (Spanien), die Johannes Kepler Universität Linz (Österreich), Integromics (Spanien), Hospital Carlos Haya (Spanien), sowie das Leibniz Rechenzentrum (Deutschland) angehören. Im Besonderen soll durch dieses Projekt die Zusammenarbeit zwischen akademischen und industriellen Partnern durch den Austausch von Personal gefördert werden. Das Projekt verbindet durch seinen interdisziplinären Zugang Life Sciences als Anwendungsbereich



unter anderem mit Techniken der Bioinformatik sowie des Cloud Computings. Dies wird durch die geplante Verarbeitung von großen Datenmengen notwendig, die durch moderne Genomsequenzierungstechniken erzeugt werden. Diese stellt auch die Hauptmotivation für die Neuentwicklung von Applikationen aus diesem Bereich dar, da existierende Software-Werkzeuge nicht auf die Verarbeitung von vollständigen Genomen ausgelegt sind.

Thematische Kernpunkte der geplanten Arbeit sind Vergleiche sowie Visualisierungen von umfangreichen Genomsequenzen bis hin zu vollständigen Genomen sowie von phylogenetischen Bäumen. Ein weiteres Ziel ist die benutzerfreundliche Aufbereitung der Ergebnisdaten durch Visualisierungtechniken für verschiedenste Endgeräte, wie Tablet-PCs oder Virtual Reality Umgebungen.

Die eingesetzten Visualisierungstechniken umfassen beispielsweise Dotplots, bei denen Genomsequenzen durch paarweises Auftragen auf Diagrammachsen sowie Markieren von äquivalenten Teilsequenzen erstellt werden. Um diese Projektziele zu erreichen, wird eine Software-Lösung bestehend aus Cloud Computing- und Hochleistungsrechenkomponenten, bioinformatischen Algorithmen, sowie Modu-

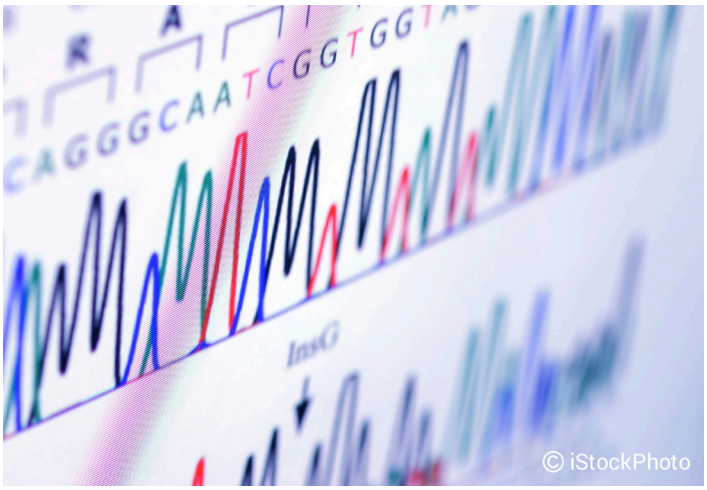
Softwarepark Hagenberg, Hauptstraße 90, A-4232 Hagenberg
T +43 736 33 43-0, office@softwarepark-hagenberg.com, www.softwarepark-hagenberg.com

SOFTWAREPARK HAGENBERG - AUS IDEEN WERDEN ERFOLGE

Big Data Analysis in Life Sciences: Cloud Computing für die Genomanalyse

len für Datenzugriff und die Visualisierung auf verschiedensten Ausgabegeräten entwickelt. Diese Module können in weiterer Folge zu komplexen Workflows zusammengesetzt werden.

Für die erfolgreiche Realisierung dieser Vorhaben bringt die RISC Software GmbH in enger Zusammenarbeit mit dem Forschungsinstitut RISC im speziellen ihre Kompetenz in den Bereichen symbolisches Rechnen sowie Cloud Computing ein und trägt in enger Kollaboration mit dem Leibniz Rechenzentrum zur Entwicklung der Cloud Computing- sowie Visualisierungskomponenten bei. Des Weiteren ist die RISC Software GmbH für die Öffentlichkeitsarbeit im Rahmen des Mr.SymBioMath Projekts hauptverantwortlich.



Als Vorarbeit für dieses Projekt wurde von der RISC Software GmbH ein auf dem Datenverarbeitungsframework Hadoop (http://hadoop.apache.org) basierendes Verfahren zum Vergleich von Genomsequenzen implementiert, um die technischen Möglichkeiten in Bezug auf Datenvorverarbeitung, Vergleichsverfahren sowie Visualisierungsmöglichkeiten auszuloten.

Verteilte Recheninfrastrukturen

Der Sequenzvergleich wurde mit Hilfe des Open-Source-Frameworks Hadoop umgesetzt, welches zur verteilten Verarbeitung von großen Datenmengen eingesetzt wird und sich daher besonders gut für die gewählte Aufgabenstellung eignet. Zu diesem Zweck werden sogenannte MapReduce-Jobs mittels Master-Worker-Prinzip automatisiert auf einem Cluster verteilt. Um diese Aufteilung möglichst effektiv nutzen zu können, werden auch die zu verarbeitenden Daten im Hadoop Distributed File System (HDFS) verteilt abgespeichert.

Da die Ergebnisse von Sequenzvergleichen mit Wörterbüchern mit mehreren Millionen von Einträgen verglichen werden können, wurde bei der Speicherung der Ergebnisse die Open-Source-Bibliothek HBase ausgewählt. Auf HDFS aufbauend bietet HBase die Möglichkeit große unstrukturierte oder semistrukturierte Datenmengen in Form von Tabellen abzuspeichern. HBase-Tabellen können auf einem Hadoop Cluster sowohl aufgebaut als auch effizient verarbeitet werden.

Visualisierung

Schließlich wurde zur Informationsgewinnung eine interaktive Visualisierungsanwendung entwickelt, welche die Ergebnisse von Sequenzvergleichen als Dotplot darstellt. Bei der finalen Anwendung ermöglicht nun eine Webservice-Schnittstelle die Auslagerung der speicher- und rechenintensiven Operationen und unterstützt damit eine vielfältige Verwendbarkeit der Anwendung auf Geräten mit wenig Speicher- und Rechenkapazität, wie Smartphones und Tablets.

Softwarepark Hagenberg, Hauptstraße 90, A-4232 Hagenberg
T +43 736 33 43-0, office@softwarepark-hagenberg.com, www.softwarepark-hagenberg.com